



UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
DEPARTAMENTO DE COMPUTACIÓN

Título de Tesis

Tesis de Licenciatura en Ciencias de la Computación

Damián Eliel Aleman

Director: Juan Manuel Pérez y Santiago Kalinowski

Codirector: Agustín Gravano

Buenos Aires, 2017

Resumen

El crecimiento de la cantidad de datos en la web y los recursos computacionales de las últimas décadas da la posibilidad de investigar fenómenos lingüísticos a gran escala, tarea casi imposible de realizar manualmente.

En el presente trabajo nos proponemos estudiar los contrastes léxicos en las distintas regiones de la Argentina a través de la creación de un conjunto de datos propio recolectado a partir de Twitter.

Índice general

1.. Introducción	1
1.1. Trabajo previo en el área	2
1.2. Twitter	2
1.3. Lingüística de Corpus y Lingüística computacional	3
2.. Datos: Extracción y procesamiento	6
2.1. Extracción de Datos	7
2.1.1. Búsquedas geolocalizadas	7
2.2. Tokenización y normalización	8
2.3. Caracterización de la muestra	9
2.3.1. Distribución temporal de tuits	10
3.. Métricas para detectar palabras salientes	11
3.1. Búsqueda de contrastes	12
3.1.1. Métricas para medir el contraste en la frecuencia de las palabras	13
3.1.2. Valor de información	13
3.1.3. Frecuencia de las palabras	15
3.2. Distribución de la entropía	16
3.3. Distribución del valor de la información	17
3.4. Proporción acumulada de ocurrencias	18
4.. Análisis de las palabras contrastivas encontradas	21
4.1. Palabras candidatas	22
4.2. Regiones de palabras	22
4.3. Problemas en el conjunto de datos	22
4.4. Caracterización de las palabras identificadas como contrastivas	24
4.5. Validación estadística	27
4.5.1. Test Hipergeométrico	27
4.5.2. Test t de Welch	29
5.. Conclusiones y trabajo futuro	32
5.1. Conclusiones y trabajo futuro	33
6.. Apéndice	34
6.1. La entropía como medida del desorden	35
6.2. Tablas y Gráficos	35
7.. Bibliografía	38

1. INTRODUCCIÓN

1.1. Trabajo previo en el área

La dialectología es un campo que estudia la variación del lenguaje según la región geográfica y el contexto social en el que se utiliza. La investigación cuantitativa en estos campo suele utilizar las frecuencias de variables lingüísticas: atributos fonéticos, sintácticos y léxicos.

Una palabra es contrastiva cuando la frecuencia de uso en dos regiones es muy diferente. Actualmente, los métodos por los cuales se descubren palabras con contraste léxico en distintas regiones consisten en cuestionarios como el de *Almeida*[AV95]. Estas encuestas están integradas por grupos temáticos centrales como la casa, la familia, la enseñanza, el cuerpo humano, etc. Sobre cada grupo temático se les indicaba a las personas entrevistadas un repertorio de palabras por cada noción, para ver si las conocían y con qué frecuencia se las usaba. Con este trabajo planteamos cambiar el paradigma y detectar automáticamente las palabras usadas en distintas regiones y sus frecuencias.

Uno de los corpus lingüísticos del español más reconocidos es el *corpus XXI*[Esp], creado por la Real Academia Española con una distribución de 25 millones de formas por cada uno de los años comprendidos en el periodo 2001 a 2012. Sin embargo, dicho corpus tiene dos desventajas importantes: por un lado, la cantidad de palabras de América Latina están subrepresentados en relación a la demografía ya que el 34,30 % de las palabras del dataset provienen de textos de España y 65,70 % de los demás países hispanoparlantes. Por otro lado, uno no dispone de todo el dataset, sino que solamente se pueden hacer las consultas desde su página web. Estas consultas están limitadas en cuanto a la cantidad de solicitudes y a las funcionalidades que estas proveen.

Una de las virtudes de hacer un corpus con un método de recolección de textos de forma automática se desprende del mayor tamaño del corpus en comparación con métodos manuales como digitalización de textos.

A pesar de haber comenzado hace varias décadas la recolección de textos de la web para realizar corpus, no hay muchos en el idioma español. Uno es el de *Mark Davies*, en el cual se utilizó las páginas web para recolectar los textos, con dos billones de palabras en español y dividió las páginas a partir del país de origen identificado por *Google*.

1.2. Twitter

Twitter¹ es un servicio de microblogging creado en 2006. Los usuarios son variados, desde personas, instituciones gubernamentales y no gubernamentales hasta bots (i.e programas que corren tareas automáticamente). Cada usuario puede escribir textos llamados tuits, que tienen una longitud máxima de 140 caracteres². Las relaciones en Twitter no necesitan ser recíprocas. Es decir, uno puede seguir a una persona en cuyo caso va a poder leer todos los tuits generados por ella, como también puede ser seguido por una persona. Un tuit puede ser respondido, como también puede ser retuiteado. El retuit es un mecanismo para diseminar por la red tuits generados por otros usuarios. De esta manera si un usuario *A* realiza un retuit generado por el usuario *B*, cualquier seguidor de *A* también va a recibir ese tuit en su panel, al cual llamaremos *timeline*. Si bien los tuits que se ven en el timeline son solo aquellos generados por los usuarios que uno sigue, todos los tuits son públicos, es decir que pueden ser accedidos a través de búsquedas en la plataforma.

¹ www.twitter.com

² Recientemente han aumentado el límite a los 280 caracteres.

Para dar una noción de la cantidad de usuarios en la Argentina, en el año 2016 había 11,8 millones de usuarios de Twitter en la Argentina. Teniendo en cuenta que en ese momento 15 millones de personas tenían smartphones se deduce que el 70 % de la gente con smartphones poseía una cuenta de Twitter.

Las ventajas de Twitter sobre otras plataformas son varias: provee una interfaz pública para obtener tuits de cualquier persona, independientemente de que haya una relación con el usuario que escribió los tuits en la red social. Es decir, uno puede ver los tuits de otra persona, sin necesidad de ser un *seguidor* de esta. Además, a diferencia de un portal de noticias donde los comentarios suelen estar relacionados con estas, en Twitter son más amplios los tópicos de los comentarios. Por otro lado, Twitter es una tecnología que permite hacer escalable el trabajo a diferentes países ya que con una misma interfaz se pueden obtener los textos de cualquier región. En cambio, si se elige un portal de noticias para sacar comentarios de usuarios, es necesario conocer la estructura de cada página para obtener esos datos. Otra virtud de Twitter es la identificación de los usuarios sobre la cual se podrían llegar a inferir datos como género, edad y ubicación de cada uno. Por último, una ventaja sobre otro método de recolección es que a través de Twitter se pueden recolectar textos con una granularidad regional muy variable y obtener información de quienes los escriben. Existen otras redes sociales de las que se podrían obtener textos para analizar, pero tienen la desventaja de ser privadas, como *Facebook*, o acotadas en términos de los temas que se hablan, como *Linkedin*.

En los últimos años se han publicado numerosos trabajos que utilizaron datos de Twitter, desde la detección y monitoreo de terremotos en tiempo real [SOM10], análisis de sentimientos y de la opinión pública [Liu12], predecir el mercado de valores [PP10] o los resultados de elecciones nacionales [TSSW10]. También se ha utilizado para localizar enfermedades por región [PD11].

En cuanto a trabajos relacionados con la lingüística cabe mencionar el trabajo de Eisenstein et al. [EOSX10] en el que identifica palabras con una gran afinidad regional realizando un modelo probabilístico en el que asumen que las distribuciones léxicas dependen de la región geográfica y de una división de tópicos. En otras palabras, suponen que hay una división de temas sobre todo el dataset y dependiendo de la región del autor, este es más propenso a escribir con una variación dada. El mismo autor realizó un trabajo para identificar variables léxicas y detectar regiones dialectales [Eis14]. El trabajo de Gonçalves et al. [GS14] consistió en analizar las variaciones diatópicas de ciertos conceptos en las grandes ciudades hispanoparlantes. Utilizaron la técnica K-means [Bis06] para obtener regiones dialectales. Por otro lado G. Doyle et al. [Doy14] propone un método bayesiano para estimar la distribución de la frecuencia de una palabra (o frase) condicional a la ubicación de la persona que la escribe.

1.3. Lingüística de Corpus y Lingüística computacional

La lingüística de corpus es una área que utiliza una serie de procedimientos o métodos para estudiar el lenguaje [MH11]. Esta rama de la lingüística intenta responder preguntas asociadas a la utilización del lenguaje a través de conjuntos de muestras de uso de la lengua. Aunque lo más común es que estas muestras provengan de textos, también se puede extraer datos a partir de grabaciones de voz o videos. Si bien esta rama nació realizando y analizando corpus de forma manual, el rápido crecimiento tecnológico de las últimas décadas dio la posibilidad de tener corpus con millones de palabras y realizar algunos

estudios de manera más automatizada, utilizando menos recursos humanos y ahorrando tiempo. A continuación, haremos un breve resumen de la lingüística de corpus.

En el año 1967 se publicó el primer corpus con un millón de palabras denominado Brown Corpus, siendo así uno de los pioneros en la lingüística de corpus. Recién en el año 1995 se consiguió realizar un corpus del inglés británico con 100 millones de palabras, titulado British National Corpus (BNC). Este corpus se originó con el objetivo de ser una muestra representativa del inglés británico de aquella época. Es importante destacar, por la gran cantidad de recursos que se utilizaron, que este trabajo se hizo con la colaboración de tres grandes editoriales, la Universidad de Oxford, la Universidad de Lancaster y la Biblioteca Británica. El 90 % del corpus era de origen escrito y el 10 % restante sobre grabaciones de conversaciones transcritas, de voluntarios de distintas edades, clases sociales y regiones. Estas conversaciones fueron producidas a partir de diferentes situaciones, algunas formales como reuniones de gobierno y otras más informales como programas de radio. Una de las grandes diferencias entre el BNC y los corpus ya existentes en ese momento, es que además de publicar los datos para investigaciones académicas, sino que también se dio acceso a los datos para uso comercial y educativo.

El crecimiento de la cantidad de datos generados mediante sistemas informáticos en las últimas décadas fue tal que en el año 2003 Kilgariff et al. [KG03] se preguntaron acerca de la posibilidad de utilizar la Web como fuente para recolectar textos. La Web resulta de una gran oportunidad para el estudio de las lenguas ya que provee una cantidad inmensa de datos, accesibles de forma gratuita y con disponibilidad inmediata. Hay varias críticas que se le pueden hacer al contenido que se encuentra en la web, como los errores sintácticos y gramaticales. Sin embargo, la inmensa cantidad de datos que es posible recolectar ofrece una oportunidad única para realizar estudios lingüísticos.

En particular, la lengua utilizada en las redes sociales nos brinda la posibilidad de identificar palabras muy asentadas en determinada región del español, difícil de detectar manualmente en la literatura. Esta dificultad proviene de varios factores. Por un lado, encontrar gran cantidad de autores nativos de diferentes lugares no es una tarea sencilla. Por otro lado, en la literatura se suele utilizar un vocabulario más restringido, normalmente excluyendo (o utilizando con menos frecuencia) términos del habla cotidiana. Un ejemplo de esto son los coloquialismos, cuyo uso es notablemente más frecuente en las redes sociales que en la literatura.

La gran importancia de saber el uso de las palabras en ciertas regiones se puede ver reflejado en las marcas geográficas (o diatópicas) que se encuentran en algunas entradas de los diccionarios. Esta información cobra importancia para saber, por ejemplo, si una palabra tiene un uso general o se la utiliza comúnmente en algunas regiones. El área de la lingüística que estudia los principios teóricos en que se basa la composición de diccionarios se conoce como lexicografía. Históricamente se han hecho diccionarios hispanoamericanos comparando con el español que los diccionarios españoles, generalmente con el diccionario de la Real Academia Española (Diccionario de la Lengua Española) considera general (ver [Zim06]). Esto ocurrió en parte por el gran desarrollo de los diccionarios de la lengua española a principios del siglo XVIII y por el carácter incipiente de la lexicografía americana. Sin embargo, esta metodología que compara dialectos de países latinoamericanos con España tuvo su rechazo en las últimas décadas, especialmente porque «una comparación adecuada es la que se puede establecer entre los elementos de entidades equivalentes, como las que forman los países.» ([Ávi04]). Tanto Raúl Ávila como Klaus Zimmermann ponen en discusión el sentido de los diccionarios diferenciales de cada país, principalmente por

no ser autosuficientes, ya que un diccionario diferencial no se encuentran las palabras que se usan en ambas regiones, sino que aparecen únicamente los términos cuyo uso es mayor en la región a estudiar sobre la región con la que se compara. Ambos autores concluyen que es de mayor interés realizar diccionarios integrales, donde se marquen las palabras que se usan de forma contrastiva en una región, pero que también estén las palabras de uso general. Creemos que la metodología propuesta en esta tesis, facilitará el armado de estos diccionarios en particular y el estudio del léxico español hispanoamericano en general.

[Baa01]

En este trabajo presentamos un método semi-supervisado para la detección de palabras contrastivas a través de un conjunto de textos recolectados de Twitter. Si bien se recolectaron textos de la Argentina, este trabajo puede ser replicado sobre otras regiones. Cabe mencionar que la herramienta detecta palabras con valores significativos de contraste en su uso, es necesaria la supervisión de investigadores lexicógrafos entrenados para seleccionar los términos con interés lingüístico.

En la sección 2 explicaremos la metodología para extraer los datos de Twitter y presentamos la caracterización de la muestra. Luego en la sección 3 se muestran la métricas creadas para medir la contrastividad de una palabra y el análisis de estas .

En la sección 4 mostramos las palabras identificadas como más contrastivas a partir de la métrica elegida y la proporción acumulada de sus ocurrencias en regiones de pocas provincias. También detallamos la validación lingüística realizada por la Academia Argentina de Letras, exhibimos una caracterización de las palabras salientes y hacemos una validación estadística de la métrica a través de tests estadísticos.

Finalmente en la sección 5 sacamos conclusiones a partir de los resultados obtenidos e indicamos trabajos posibles para seguir la investigación.

2. DATOS: EXTRACCIÓN Y PROCESAMIENTO

2.1. Extracción de Datos

Para la recolección de tuits, primero se extrajo una cantidad de usuarios con información geográfica disponible con el fin de obtener todos sus tuits. Los usuarios se buscaron por provincia de modo tal que haya una cantidad aproximadamente equitativa de cada una. La búsqueda de los usuarios se hizo de la siguiente manera:

Por cada provincia de la Argentina, se extrajo las ubicaciones de cada uno de sus departamentos, de los partidos de la provincia de Buenos Aires y de las comunas de la Ciudad Autónoma de Buenos Aires. El conjunto de estas forman la subdivisión de segundo orden de la república Argentina. La lista de departamentos/partidos/comunas fue extraída a partir de los datos publicados del Censo Argentino del año 2010. Para extraer los tuits se utilizó la librería de *python* llamada *tweepy*. De esta manera se recolectó aproximadamente 2000 usuarios por provincia, lo que resulta en 46000 usuarios argentinos. Sobre este conjunto de usuarios se buscaron los tuits. Se decidió no tener en cuenta los retuits dado que no son escritos por los usuarios sino que son una mera copia de otros tuits.

2.1.1. Búsquedas geolocalizadas

Las búsqueda geolocalizada es una herramienta provista por la API (*Application Programming Interface*) de *Twitter* en la cual primero se intenta de buscar tuits cuyas coordenadas sean las buscadas. En caso de no tener éxito, se busca aquellos tuits creados por usuarios que tienen en el campo *location* de su perfil un lugar cuyo geocódigo coincida con el de sus coordenadas. Es decir, si se hace una búsqueda inversa de las coordenadas, devuelve el lugar de su perfil.

Una vez obtenida la lista de ubicaciones, se realizaron búsquedas por cada provincia con centro en las coordenadas de los departamentos de la misma y con un radio de 20 millas. Sobre el resultado de esta búsqueda, únicamente se seleccionaron los usuarios que tienen como campo *location* al menos uno de los nombres de las ciudades de la provincia. Con esta precaución eliminamos los posibles tuits de turistas que escribieron en un lugar pero que no viven allí.

En el gráfico de la figura 2.4 se muestran las ubicaciones de los usuarios.

En la figura 2.1 se ve que la distribución de usuarios en las provincias es bastante pareja. Si bien en la figura 2.2 hay regiones grises que indican la ausencia de usuarios en ese lugar, cabe destacar que los mapas se realizaron obteniendo las coordenadas geográficas a partir de la ubicación definida en el perfil del usuario. Por lo tanto, si una persona declara que vive en *Tucumán, Argentina*, contabilizamos como que esa persona vive en la capital de esa provincia, lo cual puede no ser cierto. Sin embargo, esto no invalida los resultados puesto que la granularidad del análisis es a nivel provincial. Finalmente para ver la distribución de las coordenadas de los usuarios a lo largo del país mostramos la figura 2.3. Se puede observar que en la mayoría de las grandes ciudades hay usuarios en nuestro conjunto de datos. En el apéndice se puede encontrar un mapa 6.1 donde se contabilizaron todos los tuits con coordenadas geográficas del conjunto de datos. En este gráfico se puede observar que la distribución es mucho más amplia, aunque sigue habiendo más concentración de usuarios en aquellos departamentos con más densidad poblacional.

Si bien en este trabajo nos enfocamos en las coordenadas de las localidades dentro de Argentina, basta con cambiar las coordenadas y los nombres de las localidades que tienen

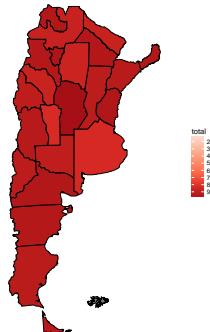


Fig. 2.1

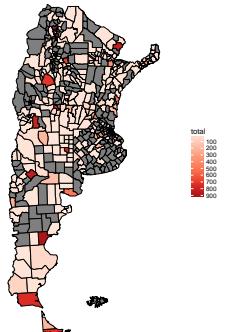


Fig. 2.2



Fig. 2.3

Fig. 2.4: Ubicaciones de los usuarios: En la figura 2.1 se muestra un mapa de Argentina con la distribución de los usuarios en las provincias sobre el conjunto de desarrollo. El mapa de la figura 2.2 permite visualizar la distribución de los usuarios en los departamentos sobre el conjunto de desarrollo. Se muestra con áreas grises los departamentos que no poseen usuarios que hayan definido el campo ubicación de su perfil en ese lugar. El mapa 2.3 muestra la distribución de las coordenadas obtenidas a partir de todos los usuarios del conjunto de desarrollo. Las coordenadas fueron obtenidas a través de un proceso de geocodificación.

que tener los campos *location* para realizar un análisis sobre otros países en una segunda etapa.

2.2. Tokenización y normalización

En cuanto al análisis del texto surge una primer problemática: ¿qué es una palabra? En principio podemos definir a una palabra como cualquier secuencia de caracteres delimitados por espacios blancos. Con esta definición 523456 y ? serían palabras. Debido a esto podríamos restringir nuestra definición a una secuencia de caracteres alfabéticos. Los ejemplos mencionados anteriormente dejarían de estar dentro de la definición. Sin embargo términos como *asdsdafsdf* también serían palabras. Para evitar este problema podríamos tener un diccionario como filtro para saber si una secuencia de caracteres dada es una palabra. Si bien esto tendría mucha precisión al momento de filtrar los términos, no sería capaz de detectar palabras que existen en una lengua pero que no están incluidos en el diccionario elegido. Además, dada la cantidad de palabras recogidas, es altamente improbable que una secuencia al azar de caracteres alfabéticos reúna las condiciones de frecuencia necesarias para resultar destacada por la métrica que utilizamos. Es por eso que decidimos tomar a una palabra como una secuencia de caracteres alfabéticos.

Es muy posible que tengamos palabras que no sean interesantes a nivel lingüístico, como errores de tipeo (e.g computadira, escribur), errores ortográficos o nombres propios. Es importante destacar que Twitter tiene caracteres especiales para mencionar a la gente, como el @, o el #(hashtag) utilizado para agrupar mensajes. Estos caracteres aparecen mucho, ya que los usuarios suelen responderse en la red, mencionando los mismos temas (aclarando el hashtag), o respondiendo a otros usuarios. Ya que esos caracteres no son alfabéticos, cualquier término que los utilice no va a ser parte del conjunto de palabras, como tampoco lo serán las direcciones de páginas web. Decidimos que se filtren estos términos ya que no tienen interés lingüístico y además agregarían mucho ruido a los datos.

Además de la tokenización del texto, se realizó una normalización sobre él. Todas las letras se convirtieron a letra minúscula y las palabras con más de tres letras iguales de forma consecutiva se redujeron para que solo tengan tres repeticiones. De esta forma, el término *padreeeee* y *padreeee* fueron reducidos a una única unidad léxica (*padreee*). Esto se hizo con la librería *TweetTokenizer* de *NLTK*. Se descartó la idea de filtrar las palabras que no estuvieran en un diccionario ya que si bien hubiera eliminado mucho ruido, también nos hubiera filtrado palabras de interés. Este es el caso de los neologismos, o las palabras que, si bien se utilizan hace mucho tiempo, no están en los diccionarios actuales.

2.3. Caracterización de la muestra

Para tener una noción más completa de la muestra, se presenta la tabla 2.1 que indica las cantidades de palabras y tuits por provincia.

Provincia	#Palabras Distintas	#Usuarios	#Tuits	#Total Palabras
Buenos Aires	191919	920	1125042	8974372
Catamarca	173104	957	1057019	8161309
Chaco	169476	964	976943	7605991
Chubut	182592	954	1023373	8884745
Córdoba	207307	987	1224266	10075932
Corrientes	183292	939	1044951	8426940
Entre Ríos	188679	969	1193693	9462986
Formosa	169254	903	923352	7184382
Jujuy	171064	971	678004	5951778
La Pampa	186593	935	1085757	8996318
La Rioja	186041	946	704044	6757277
Mendoza	193708	945	1099717	9402399
Misiones	168400	972	984218	7790197
Neuquén	188038	927	1111201	9021449
Río Negro	194383	965	1215361	9991831
Salta	188402	884	830916	7506652
San Juan	183546	926	1002322	8377792
San Luis	164185	896	1006464	8327093
Santa Cruz	174089	935	876621	7432923
Santa Fe	201879	937	1019620	8862328
Santiago del Estero	166540	887	944109	7355729
Tierra del Fuego	197273	964	976426	8559218
Tucumán	195643	962	1093874	9238526

Tab. 2.1: Cantidades del conjunto de datos

También analizamos la cantidad de palabras por tuit promediadas sobre cada usuario. Debido a que los tuits están limitados a 140 caracteres, era de esperar que no hubiera demasiadas palabras promedio por cada tuit. En la figura 2.5 podemos observar que la media para la cantidad de palabras promedio en un tuit está entre 7 y 8.

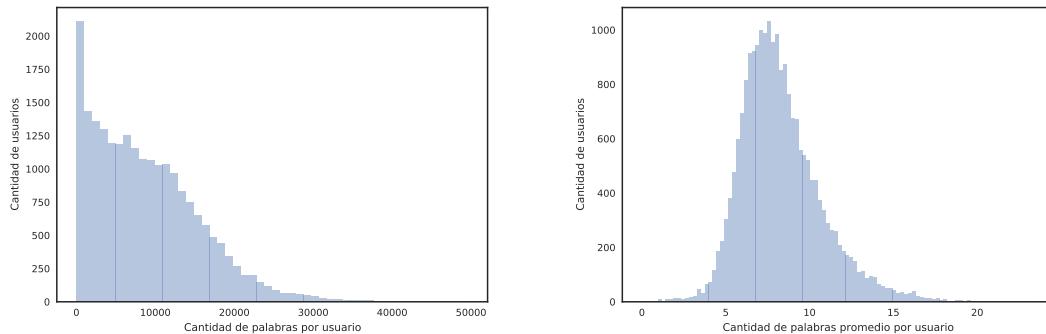


Fig. 2.5: Histograma de la cantidad de palabras totales por cada usuario.

Fig. 2.6: Histograma de la cantidad de palabras promedio para todos los usuarios.

2.3.1. Distribución temporal de tuits

Los tuits recolectados para el conjunto de datos de desarrollo tienen una particularidad: a medida que pasan los años hubo mayor cantidad de tuits durante un año. Esto se refleja en los gráficos de las figuras 2.7 y 2.8.

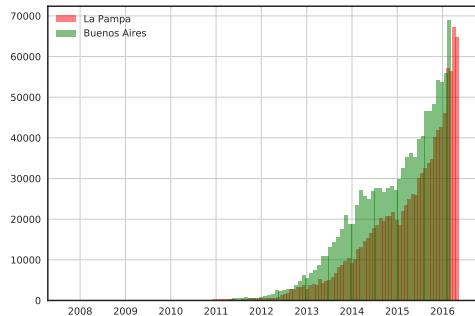


Fig. 2.7

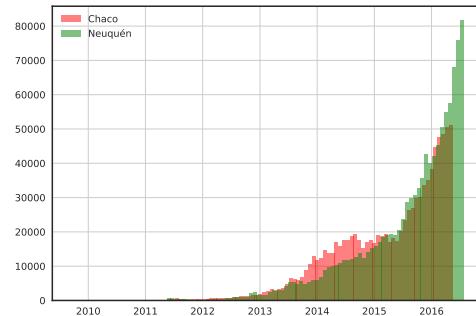


Fig. 2.8

Fig. 2.9: En la figura 2.7 se presenta un histograma donde se muestra la cantidad de tuits que se hicieron por intervalo de tiempo en las provincias La Pampa y Buenos Aires. En la figura 2.8, se presenta el gráfico para Chaco y Neuquén.

3. MÉTRICAS PARA DETECTAR PALABRAS SALIENTES

3.1. Búsqueda de contrastes

Una palabra tiene un contraste cuando esta tiene un uso con diferencias significativas en distintas regiones. En este trabajo nos propusimos crear un listado con palabras con contrastes que tengan importancia a nivel lingüístico. En este sentido, los nombres de personas, lugares u organizaciones no fueron considerados de interés a pesar de tener contrastes en su uso. Este listado fue ordenado por una métrica que capte en un único valor el nivel contrastivo. De esta manera, se seleccionó un subconjunto de palabras, de acuerdo a la métrica, el cual fue analizado manualmente en otros textos por la Academia Argentina de Letras.

El primer acercamiento para ver el contraste de las palabras lo realizamos comparando las frecuencias de las palabras en cada par de provincias de la Argentina. Para esto calculamos, por cada palabra, la frecuencia de ocurrencias sobre cada una de las dos provincias. La mayor frecuencia de ambas, la llamamos frecuencia máxima y a la menor, la frecuencia mínima. Luego el cociente entre la frecuencia máxima y la frecuencia mínima tiene como resultado lo que llamamos *maxDif*. En caso de que en una de las dos provincias no se haya recolectado tuits con esa palabra, se tomaba como frecuencia mínima a la frecuencia mínima distinta de 0 de todas las palabras generadas en esa provincia. Así se evitó la división por cero. Esta métrica se resumen en la ecuación 3.1.

$$\text{maxDif}(w, p_1, p_2) = \frac{F_{\max}(w, p_1, p_2)}{F_{\min}(w, p_1, p_2)} \quad (3.1)$$

donde

$$F_{\max}(w) = \max(frec(w, p_1), freq(w, p_2)) \quad (3.2)$$

$$F_{\min}(w) = \begin{cases} \min(freq(w, p_1), freq(w, p_2)) & \text{si } freq(w, p_1) * freq(w, p_2) > 0 \\ \min(freq(w, p)) & \forall w \in \text{palabras}(p), \text{ con } p = \{p_1, p_2\} \setminus \{P_{\max}\} \text{ sino} \end{cases} \quad (3.3)$$

donde P_{\max} es la provincia que tiene la mayor frecuencia de ambas.

De esa manera se ordenó el listado de cada par de provincias teniendo en cuenta la división de frecuencias. Sin embargo, este método imposibilitaba el trabajo manual para la Academia Argentina de Letras que debía mirar estos listados y hacer un análisis más exhaustivo sobre las palabras con mayor diferencia de frecuencias, debido a que había $\binom{23}{2} = 253$ listados (o equivalentemente 253 columnas en un mismo listado) a analizar. Además la métrica solo permitía saber si había un contraste entre dos provincias, pero no se podía tener en cuenta la frecuencia de la palabra en el resto de las provincias. En consecuencia las palabras se encontraban repetidas en los distintos listados y con diferentes valores de *maxDif*, lo cual hacía muy difícil poder identificar en qué regiones había una diferencia significativa de frecuencias.

Debido a esto decidimos realizar un nuevo enfoque para encontrar las palabras con alta contrastividad en las distintas regiones, de manera que una métrica pueda reflejar el nivel de contrastividad de la palabra en un único valor. De este modo, nos enfocamos en analizar el contraste de frecuencias de palabras sobre las provincias a través de una métrica superadora.

3.1.1. Métricas para medir el contraste en la frecuencia de las palabras

Dado que se quieren encontrar las palabras con contrastes significativos en distintas regiones se propone generar una métrica basada en la cantidad de información para poder realizar esta tarea.

Una medida que se puede usar para comparar las frecuencias de las palabras en las diferentes regiones del país puede ser la entropía definida por Shannon (ver en el apéndice: [6.1](#)), debido a que nos brinda un valor que informe qué tan uniforme es la distribución de las frecuencias de cada palabra. La entropía es máxima cuando la probabilidad de los eventos es equiprobable y mínima en el caso que la probabilidad de un evento es 1. Sin embargo, la entropía como única medida tiene sus desventajas. En particular, una palabra con una sola ocurrencia en una provincia y ninguna en las demás, tiene la entropía mínima. A pesar de que nos interesan las palabras con un contraste significativo entre regiones, dentro de ellas elegiremos las que tienen mayor cantidad de ocurrencias. Es por esto que elaboramos otra métrica que tenga en cuenta la entropía, entre otras variables a tener en cuenta.

3.1.2. Valor de información

La métrica que utilizamos para ordenar los listados de palabras y detectar cuáles son las que tienen altos contrastes en su uso en distintas regiones fue inspirada por el trabajo de Zanette y Montemurro [[MZ10](#)]. Ellos, a diferencia de Shannon, estudiaron una relación entre una medida de la información y su función semántica en el lenguaje. A continuación detallamos el procedimiento para calcular lo que ellos llamaron el valor de la información:

Dado un texto dividido en P partes iguales llamadas ventanas, se calcula la entropía $H(w)$ sobre el vector de cantidad de ocurrencias en cada una de las P ventanas. Luego se define $\widehat{H}(w)$ como la entropía de una permutación aleatoria del texto y promediada por todos las posibles realizaciones de la permutación de él.

Es decir, se distribuyen uniformemente las palabras en P partes y se calcula la entropía como se hizo con el texto original. Es de esperar que en la mayoría de casos la entropía del texto permutado sea mayor que la medida en el cálculo original. Esto se debe a que las palabras se distribuyen de forma más uniforme en las distintas partes.

Finalmente, definen al valor de la información como $I(w) = p(w)(\widehat{H}(w) - H(w))$, con $p(w)$ la frecuencia total de la palabra en el texto. De esta manera se les da más importancia a las palabras que son más frecuentes y a las palabras que tienen una baja entropía, ya que en estas el término de la diferencia es más grande.

Este estudio se hizo sobre tres textos, *Análisis de la mente*, *Moby Dick* y *El origen de las especies* de Charles Darwin. En los tres libros las palabras con mayor valor de la información están altamente relacionadas con los temas principales.

Si bien esta métrica tiene en cuenta la frecuencia de las palabras además de la entropía, el texto en Twitter resulta difícil de dividir en partes iguales. Esto es porque la división está pensada para dividir el texto en secciones que posiblemente hablen de distintos temas y nuestros textos son tuits que por lo general no superan las 10 palabras. Otra dificultad que surge de esta métrica es la imposibilidad de realizar la media de todas las posibles permutaciones del texto por la limitación computacional ya que tenemos una cantidad muy grande de datos. Es por eso que realizamos una métrica parecida.

Podemos pensar a las palabras del texto como una variable aleatoria W , donde cada palabra w tiene una probabilidad de aparición en una provincia dada de la Argentina.

Esta probabilidad la aproximamos con la frecuencia en la que aparece, es decir la cantidad de ocurrencias de la palabra dividida por la cantidad de palabras totales. Por otro lado sea P una variable aleatoria que cuenta la cantidad de personas que utilizan la palabra p en cada provincia.

Luego, sea $cant_w(p)$ igual al logaritmo sobre la cantidad de ocurrencias de esa palabra en toda la Argentina, es decir $cant_w(p) = \log_2(cantidadOcurrencias(p))$. y sean las constantes MIN_w y MAX_w definidas de la siguiente manera:

$$MIN_W = \min_{p \in Palabras} cant_w(w) \quad (3.4)$$

$$MAX_W = \max_{p \in Palabras} cant_w(w) \quad (3.5)$$

Realizamos una normalización lineal de la función $cant_w$,

$$norm_w(p) = \frac{cant_w(p) - MIN_W}{MAX_W - MIN_W} \quad (3.6)$$

De esta manera, $norm_w$ tiene su imagen en el rango $[0, 1]$, tomando el valor 0 sobre la palabra que tiene la cantidad de ocurrencias máxima y toma el valor 1 cuando se aplica a la palabra con menor cantidad de ocurrencias. Vale la pena aclarar, que tomamos 40 como umbral mínimo de cantidad de ocurrencias de las palabras para ser estudiadas. A partir de la función $norm_w$ definimos el valor de la información de las palabras I_w como:

$$I_w(w) = norm_w(w) * (\hat{H}_w(w) - H_w(w)) \quad (3.7)$$

siendo $H_w(w)$ la función de entropía calculada sobre las cantidades de ocurrencias de la palabra w sobre las 23 provincias argentinas. De forma similar \hat{H} es la función de entropía sobre las cantidades de ocurrencias simuladas en todas las provincias a partir de una distribución multinomial. Elegimos esta distribución ya que con esta se distribuye la suma de los valores de la variable aleatoria, en nuestro caso la cantidad de ocurrencias de la palabra w , de forma uniforme.

Ahora bien, una determinada provincia o región pueden tener muchas ocurrencias de una palabra formuladas por algunos pocos usuarios que utilizan constantemente el término. Un ejemplo de esto podrían ser bots que escriben automáticamente textos iguales (o similares) en grandes cantidades. Otra posible causa de este fenómeno podría ser la de usuarios que hablan de personas, lugares o marcas de forma constante. Es por esto que realizamos una métrica similar que tenga en cuenta la diferencia de la entropía sobre la cantidad de personas que utilizan la palabra . Agregandole el término $norm_u$, que es la constante normalizadora de la cantidad de usuarios descripta en la ecuación 3.9 obtenemos el valor de la información de las personas I_u ,

$$I_u(w) = norm_u(w) * (\hat{H}_u(u) - H_u(w)) \quad (3.8)$$

Donde,

$$norm_u(w) = \frac{cant_u(w) - MIN_U}{MAX_U - MIN_U} \quad (3.9)$$

$$\text{donde: } MIN_U = \min_{p \in Palabras} cant_u(p) \quad (3.10)$$

$$MAX_U = \max_{p \in Palabras} cant_u(p) \quad (3.11)$$

y $cnt_u(p)$ es el logaritmo sobre la cantidad de usuarios que utilizan dicha palabra en la Argentina, es decir $cnt_u(p) = \log_2(cantidadU_{\text{usuarios}}(p))$.

Debido a que queremos tener en cuenta tanto a la variación de la cantidad de ocurrencias de la palabra como a la variación de la cantidad de usuarios, elegimos como métrica la multiplicación de ambas métricas definidas, es decir

$$I(w) = I_w(w) * I_u(w) \quad (3.12)$$

Es importante aclarar que tanto $norm_w$ como $norm_u$ realizan una normalización del logaritmo de esas variables. Esto se debe a que el logaritmo genera una dispersión tal que agrupa los valores altos que se encontraban muy dispersos mientras que separa los valores pequeños que estaban concentrados. Esto se puede ver en las figuras 3.3 y 3.4.

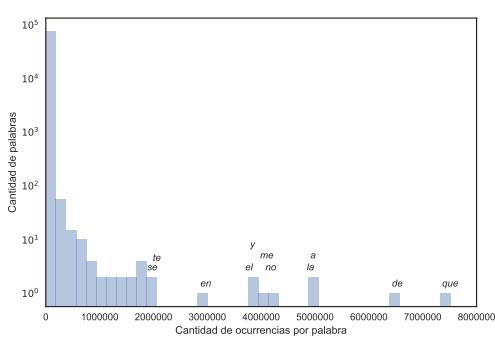


Fig. 3.1: Histograma de la cantidad de ocurrencias de las palabras.

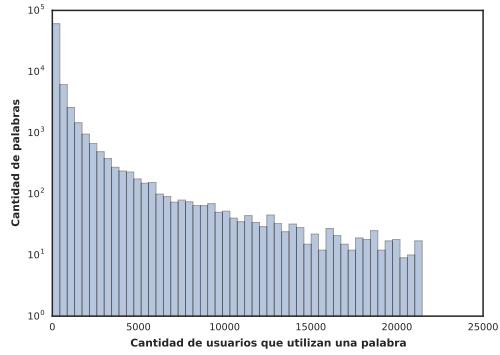


Fig. 3.2: Histograma de la cantidad de usuarios que utilizan una determinada cantidad de palabras.

Para eliminar los valores atípicos se procedió a remover tanto las palabras que no superaran las 40 ocurrencias, como también aquellas que eran dichas por menos de 6 usuarios. La métrica se evaluó en este conjunto filtrado de palabras.

3.1.3. Frecuencia de las palabras

En la figura 3.1 graficamos la distribución de la cantidad de ocurrencias de las palabras. Podemos observar que la mayoría de las palabras ocurren poco. En particular el 50 % de las palabras ocurren menos de 139 veces. Por otro lado hay pocas palabras que ocurren mucho, por ejemplo la palabra *que* o la preposición *de*.

Si comparamos la posición de la palabra en un listado ordenado podemos ver que las cantidades de ocurrencias parecieran seguir una distribución zipfiana. La ley de Zipf es una ley empírica formulada por George Zipf en el año 1932 en la cual se establece una relación entre la frecuencia de una palabra con su posición dentro del listado de palabras ordenadas por frecuencia decreciente [Mon01, Zip16]. En particular, sea n la posición de la palabra en el listado ordenado y sea $f(n)$ la cantidad de ocurrencias de la n -ésima palabra, se puede hacer la siguiente aproximación:

$$f(n) \approx \frac{A}{n^\alpha}$$

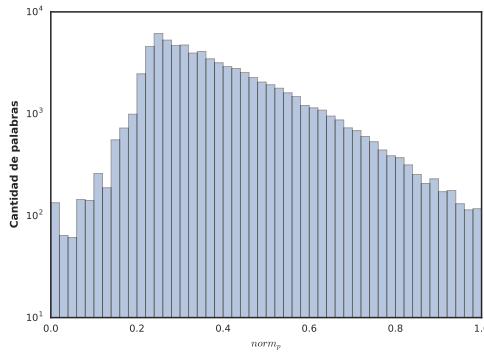


Fig. 3.3: Histograma de la cantidad de ocurrencias de las palabras.

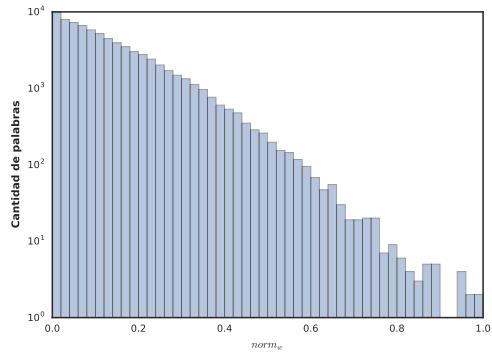


Fig. 3.4: Histograma de la distribución de cantidad de usuarios que utilizan una determinada cantidad de palabras.

donde α toma un valor levemente mayor a 1 y A es una constante normalizadora. Entonces, bajo la ley de Zipf uno puede saber que la frecuencia de la segunda palabra más dicha en un corpus, es aproximadamente la mitad que la primera. La palabra con posición 3 en el listado ordenado por frecuencias, va a tener aproximadamente la tercera parte de la cantidad de ocurrencias que la primera y así sucesivamente. De esta manera hay una relación lineal entre el logaritmo de la posición del listado ordenado por frecuencias y el logaritmo de la cantidad de ocurrencias de cada palabra. Otra forma de utilizar esta ley empírica es la siguiente: sabiendo la posición de una palabra w en el listado ordenado por frecuencias de un corpus **A** y sabiendo la cantidad de palabras totales de un corpus **B**, puede estimarse la cantidad de ocurrencias de w en el corpus **B**.

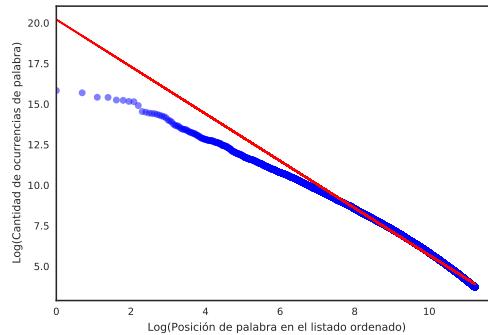


Fig. 3.5: Cantidad de Ocurrencias de palabra vs posición en listado ordenado. Se aplicó el logaritmo a las cantidades de ocurrencias, como también a los valores de las posiciones para mostrar la proporcionalidad entre $f(n)$ y $\frac{1}{n^\alpha}$.

3.2. Distribución de la entropía

Teniendo el listado de palabras hicimos un cálculo de entropía tomando en cada provincia la cantidad de ocurrencias de cada palabra. En la figura 3.6 podemos observar la distribución del valor de la entropía sobre todas las cantidades de ocurrencias de las

Palabra	Cantidad de Ocurrencias
que	7509160
de	6527014
a	4962492
la	4913854
no	4177810
me	4101998
y	3838370
el	3773455
en	2969783
te	2060662
se	1976027
un	1863075
es	1825892
con	1799979
lo	1712189
mi	1643777
por	1553382
los	1498941
para	1398757
las	1212452

Tab. 3.1: Cantidad de apariciones de las 20 palabras más frecuentes.

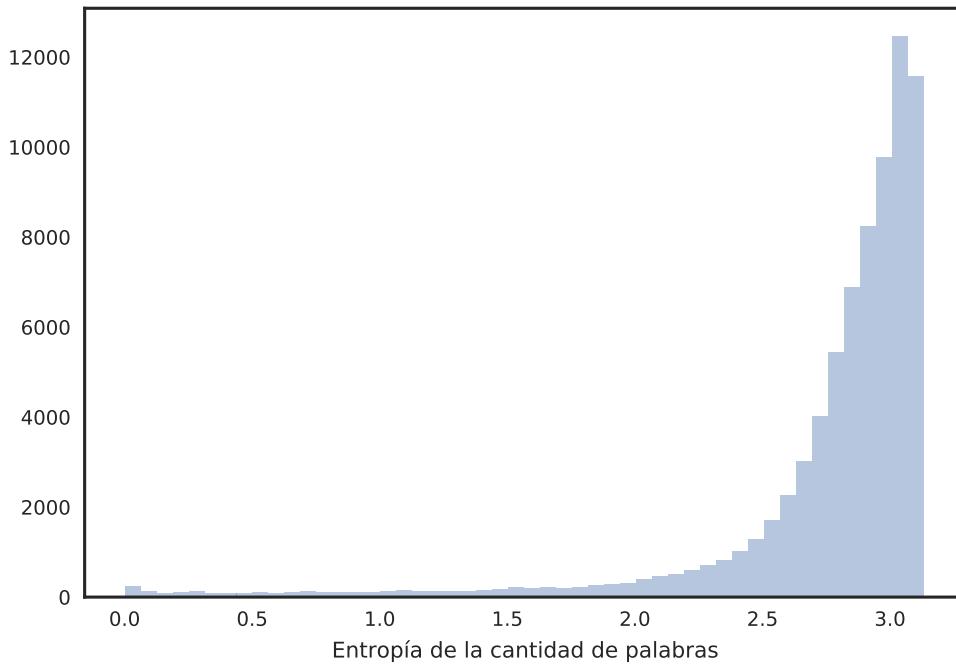
palabras con más de 40 apariciones y dichas por más de 5 usuarios.

Podemos ver que la mayor parte de las palabras tienen un valor de entropía entre 2.5 y 3. Esto quiere decir que hay un gran conjunto de palabras que tiene una cantidad de ocurrencias relativamente uniforme a lo largo de todas las provincias. Sin embargo, hay otro conjunto de palabras que tienen una entropía menor a 2, la cual podemos considerar como baja. Estas últimas palabras serán las que tienen mayor interés debido a que tienen una variación marcada en cuanto a su utilización en las distintas regiones. El máximo valor alcanzado de la entropía es de 3,1350 con la palabra *el*. Como aclaración la entropía calculada se realizó con logaritmos naturales, por lo tanto el máximo valor posible es de $\ln(23) = 3,1355$ donde habría una distribución uniforme en la cantidad de ocurrencias sobre las 23 provincias argentinas.

Tener en cuenta únicamente a la entropía de las palabras nos puede generar la detección de palabras que no son de interés, ya sea porque no ocurren una cantidad significativa de veces o porque la variación de las ocurrencias en las distintas provincias se debe solamente a pocos usuarios que la utilizan mucho. Es por esto que también se calculó la entropía teniendo como variable la cantidad de personas que utilizaron cierto término en una determinada provincia.

3.3. Distribución del valor de la información

En el gráfico 3.7 se muestra una clara relación entre la cantidad de ocurrencias que tiene una palabra y su valor de la información, indicado por el color: cuanto más oscuro



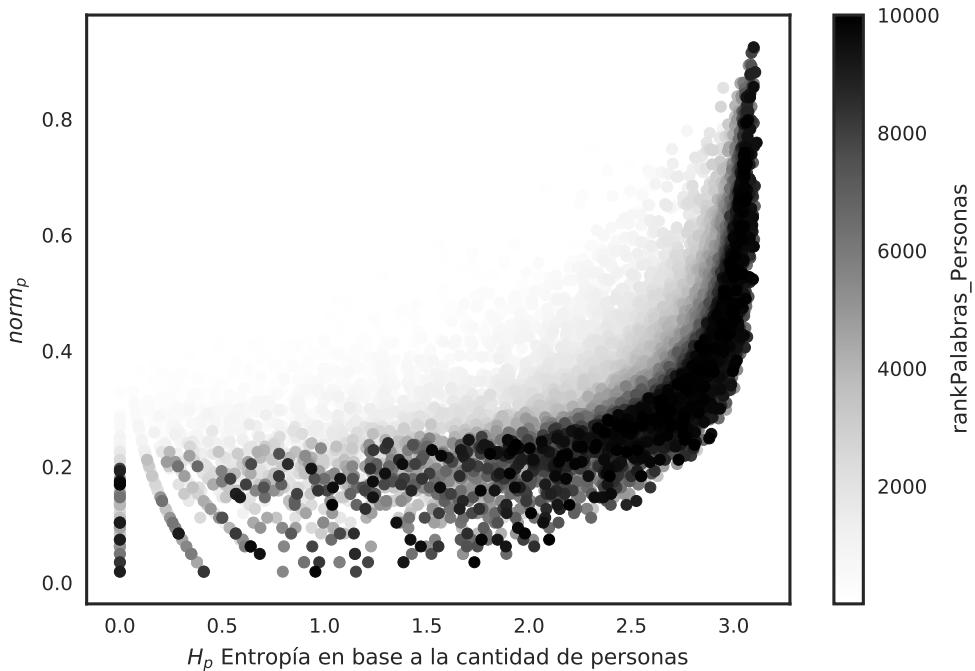


Fig. 3.7: Gráfico de dispersión que muestra la posición en el listado ordenado según el valor de la información a partir de la escala cromática. Las posiciones más bajas aparecen más blancas. A su vez se muestra para cada palabra el valor de la entropía de las personas (H_u) y la cantidad normalizada de personas que utiliza dicho término($norm_p$).

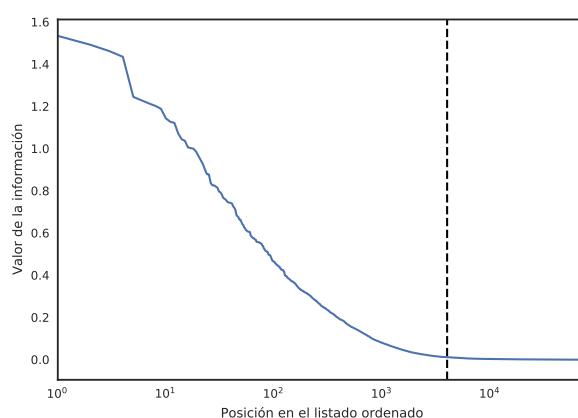


Fig. 3.8: Distribución del valor de la información según la posición de la palabra en el listado de palabras. El gráfico se realizó sobre el conjunto de palabras cuya cantidad de ocurrencias era mayor a 40 y la cantidad de usuarios que utilizaron cada término era mayor a 5.

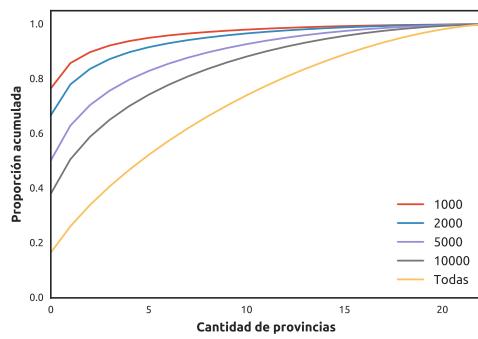


Fig. 3.9: Proporción de ocurrencias acumulada según la muestra de palabras.

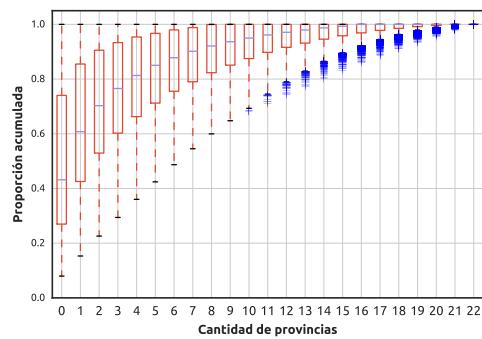


Fig. 3.10: Variación de la proporción de ocurrencias acumulada a partir de la muestra con las primeras 5000 palabras con mayor valor de la información.

4. ANÁLISIS DE LAS PALABRAS CONTRASTIVAS ENCONTRADAS

4.1. Palabras candidatas

Para buscar las palabras candidatas a tener contrastes significativos en cuanto a la cantidad de ocurrencias en distintas provincias, elegimos el conjunto de las primeras cinco mil (5000) palabras con mayor valor de la información. El número 5000 surgió de ver la distribución de los valores de la información graficado en la figura 3.8, donde hay una caída pronunciada de la métrica y a partir de la palabra cuya posición es 4000 se observa que empieza a estabilizarse con valores muy cercanos a 0. Es por esto que nos pareció razonable dar un margen de 5000 palabras para evaluar manualmente las palabras del listado y, entre estas, seleccionar las palabras con contrastes significativos que tienen interés a nivel lingüístico.

Como era de esperar los topónimos, como los nombres de ciudades y provincias, son palabras que ocurren mayormente en sus respectivas regiones. Esto causa que haya una gran variación en la cantidad de ocurrencias sobre las distintas provincias, lo que genera un valor alto en la métrica de valor de la información. Para facilitar la detección de palabras contrastivas con mayor interés lingüístico buscamos un conjunto de datos con los nombres de las localidades y departamentos de la República Argentina de modo tal que podamos resaltarlas para que el equipo de filólogos tenga una primera alerta sobre posible toponimia.

4.2. Regiones de palabras

Una vez que calculamos las regiones que cubren un umbral para cada palabra, nos propusimos analizar cuales son las más frecuentes. Para eso generamos una lista con los conjuntos de provincias cuya cantidad sea menor a 7 y los ordenamos según su frecuencia. En la tabla 4.1 se muestran los primeros conjuntos de provincias obtenidos a partir de las primeras 5000 palabras con mayores contrastes de acuerdo al valor de la información. Más precisamente se muestran las 30 regiones con mayor cantidad de palabras y cuya cantidad de provincias sea mayor a uno. Analizando esta tabla pudimos notar que la mayoría de las regiones están compuestas por provincias contiguas, es decir que para cada provincia dentro de una región hay otra provincia limítrofe. Mostramos algunos de los conjuntos de palabras de cada región en la tabla 6.1 (ver apéndice).

4.3. Problemas en el conjunto de datos

Cuando vimos las palabras con mayor valor de la información, observamos que algunas palabras de la provincia de La Rioja eran provenientes de España. Analizando la causa de este problema, notamos que la API de Twitter no realiza las búsquedas localizadas como uno esperaría. En particular, no solo se fija en los tuits geolocalizados, sino que también hace una búsqueda inversa a través de los nombres de las ciudades que tienen esa coordenada. Específicamente, La Rioja es una provincia Argentina, como así también una provincia de España. Es por eso que al hacer búsquedas con las coordenadas de ciudades de La Rioja en Argentina, tuvimos resultados de tuits de España. Lo mismo sucedió con San Juan (capital de Puerto Rico), Santiago Del Estero (Santiago de Chile) y Córdoba (ciudad de Andalucía, España). A pesar de que los tuits no fueron escritos en Argentina, consideramos que su cantidad no es lo suficientemente grande como para tener resultados incorrectos.

Conjunto de provincias	Cantidad de Palabras
Jujuy - Salta	24
Mendoza - San Juan	19
Neuquén - Río Negro	18
Corrientes - Misiones	16
Chaco - Corrientes - Formosa	16
Chaco - Corrientes	16
Chubut - Santa Cruz	13
Catamarca - La Rioja	12
Santa Cruz - Tierra del Fuego	12
Corrientes - Entre Ríos - Formosa - La Rioja - Misiones	12
Formosa - Misiones	12
Corrientes - Formosa - Misiones	12
Córdoba - La Rioja	11
Catamarca - Salta - Santiago del Estero - Tucumán	11
Catamarca - Jujuy - La Rioja - Salta - Santiago del Estero - Tucumán	10
Chaco - Corrientes - Misiones	10
Chaco - Corrientes - Formosa - Misiones	9
Catamarca - Santiago del Estero - Tucumán	9
Catamarca - Tucumán	9
Salta - Tucumán	8
Catamarca - Jujuy - Salta - Santiago del Estero - Tucumán	8
Neuquén - San Juan	7
Chubut - Santa Cruz - Tierra del Fuego	7
Buenos Aires - La Pampa	7
Salta - Santiago del Estero - Tucumán	7
Buenos Aires - La Pampa - Río Negro	6
Corrientes - Formosa	6
Catamarca - Jujuy - Salta - Tucumán	6
Chaco - Corrientes - Entre Ríos - Formosa - Misiones	6
Catamarca - Santiago del Estero	6

Tab. 4.1: Indica cuantas palabras tienen un cubrimiento del 80 % de sus ocurrencias en cada conjunto de provincias a partir de las 5000 palabras con mayor contrastes (de acuerdo al valor de la información).

4.4. Caracterización de las palabras identificadas como contrastivas

Dentro de las palabras contrastivas identificadas a través de la métrica, podemos hacer una caracterización de ellas según el fenómeno lingüístico representan.

En base al listado de palabras identificados como contrastivas a partir de la métrica, se realizó una validación lingüística por lexicógrafos de la Academia Argentina de Letras. En esta, se realizó un estudio pormenorizado, palabra por palabra, en el cual los criterios seguidos para que una palabra sea relevante privilegiaron las posibilidades de que aquella forme parte del repertorio léxico de una comunidad de hablantes. Esto excluyó, como es tradicional en lingüística, nombres propios y topónimos locales, que la métrica sube a los puestos altos de las listas porque efectivamente su uso es abundante y contrastivo.

En la lista 4.4 presentamos una caracterización de las palabras. La lista apenas posee algunos ejemplos en cada categoría. Sin embargo sirve para ilustrar ejemplos de uso en el habla cotidiana de las palabras contrastivas identificadas. La lista completa arroja un resultado dentro del rango de las 300 palabras dignas de estudio por cada 5000 palabras, es decir, 1 palabra cada 17 aproximadamente. A pesar de que no existen otros proyectos que provean un término de comparación para evaluar el grado de éxito implicado en esta relación, no cabe ninguna duda de que, al menos en la detección de coloquialismos locales actualmente en uso, la herramienta plantea un verdadero punto de inflexión para la lexicografía contrastiva. Esta área del léxico es justamente la más elusiva, puesto que su impacto en cualquier medio impreso llega notablemente más tarde y, todavía más importante, en la mayoría de los casos no llega nunca. Se incluyeron como relevantes palabras que ya están incluidas en el Diccionario del Habla de los Argentinos [dL08], dado que ese hecho es una confirmación adicional de la pertinencia de la ubicación que asignó la métrica.

Las formas cuyo uso se exemplifica son las que están en negrita y solo en ellas se normalizó la tildación.

■ Coloquialismos o vulgarismos

«Perdon pero tenes que ser muy **culiado/a** para ir a mc y pedirte una ensalada» (Córdoba)

«Q **chombi** hacer un chiste y q la otra persona no se ría o no lo entienda» (Mendoza)

«Que **carnasas** poniendole rosas rojas a toda la ropa, para mi queda horrible sorry» (Neuquén)

■ Indigenismos

«Te regalo ser **mitaí** y ir a jurar la bandera con el guardapolvo caliente ese y la corbata que te ahorca todo (Del guaraní mitaí “pequeño”)» (Formosa)

«**Angá** mi negrito, esta triste (Del guaraní angá aprox. “pobre”) (Corrientes)» (Corrientes)

«Gracias tormenta **ura** por sonar como una pochoclera de chasquibums a las 3 de la mañana en mi ventana durante 50 minutos. (Valor despectivo. Del quechua ura “vulva, vagina”) » (Tucumán)

■ Gentilicios

Casildense (de Casilda), **concordiense** (de Concordia) y **obereño** (de Oberá).

■ Voces no marcadas en registro, que aluden a una realidad local

«Quiero a alguien que me diga vamos a comer **piadinas**, un panchito, un choripán, una hamburguesa lo que sea y soy feliz» (San Juan)

«**Tareferos** que reclamaban asistencia interzafra en Posadas estarían preparando una protesta para hoy en la Fiesta del Inmigrante en Oberá.» (Misiones)

«Me encantan los bohemios anti sistema que usan vans. Es como que seas ecologista y uses un cuaderno hecho con media **yunga**.» (Jujuy)

■ Voces sinónimas de otras más usuales en Buenos Aires

«Teres, **pororós** y pelis con Carlita y Flor» (Chaco)

«Ver un negro **chuño** con musculosa y gorro.. se ve que el tipo no quería pasar ni frío ni calor.» (San Juan)

«Tenía la re expectativa para este sábado y al final **trancó** todo » (Formosa)

■ Leísmo

«No te olvides de **saludarle** a tu suegro hoy» (Misiones)

«Vine a **visitarte** a mis primas y estan re colgadas, para eso me quedaba en mi casa no maaa » (Misiones)

«A **esperarle** a nahuel, que traiga los teresss » (Formosa)

■ Fusiones y acrónimos que pueden señalar pronunciación o alta frecuencia de uso

«Los sueños de la siesta me dejan **patra** » (Buenos Aires)

«Si mañana me dice q no, voy sola, necesito ver esa película en el cine siosi» (Córdoba)

■ Voces consideradas generales pero que, al aparecer en la lista, permitieron verificar su contrastividad en frecuencia de uso al menos con respecto a España

Ejemplos: **pavada**, **distrital** y **cariño**.

■ Voces sospechadas generales pero con acepción local diferente

«Mañana que alguien **atine** con parque y porrones» (Mendoza)

«**Mansas** ganas de sentarme a tomar un té con semitas» (San Juan)

«**Habilítenme** una nueva espaldaaa» (Tierra del Fuego)

«sigo **asada** por cosas que han pasado hace como dos días, que falla (Mendoza) / Que **asada** estoy, tengo la cabeza echada un lío» (San Juan)

■ Voces con una morfología propia de una región

Ejemplo: terminación aso/asa con base adjetiva.

«Creo que va a estar **malaso** lo de esta noche » (San Juan)

«estoy subiendo un mix re **chomoso** que hice anoche » (San Luis)

«Esta **locasa** esa mina para hacer eso» (Córdoba)

■ **Formas indicadoras de pronunciación usual**

«Menos mal que soy de los chetos de la carne y mañana tengo **asao** todo el dia jajajajaj» (Tucumán)

«Un lunes con buen humor ta **pasao** » (Catamarca)

«Ahora a la mañana tengo q ir hacerme la tarjebus jajajajj **mavale** q me estoy por levantarrr jajajaj» (Corrientes)

■ **Formas verbales coloquiales con sustantivos o adjetivos como base**

«Me calma mucho **mimosear** a mi perro » (Neuquén)

«Me vine a acostar y ya me dicen que parezco de 80 años ME CHUPA UN HUEVO LO QUE PIENSEN, DEJENME **ABUELEAR** » (Buenos Aires)

«Estaría bueno que ari venga aunque sea a saludarme y que no se quede todo el tiempo **pollereando.**» (Tierra del Fuego)

■ **Variantes ortográficas, operativas para incorporarlas algunas como tales y también para verificar la alta frecuencia de uso**

Ejemplos: culiado (adj. despect. o fórmula de tratamiento de confianza) y tereré.

«Q paja volver al colegio **culiaa**» (Córdoba)

«Que pajero el **qliao** este.» (Córdoba)

«Quiero recitaaaal **qliaaaa**» (Córdoba)

«Tereresss y pile con todos mis primisss» (Entre Ríos)

«No se si hacerme un **tere** o un mate para pasar la siesta» (Corrientes)

«Es lo mas lindo no ir al colegio y quedarme a tomar **teresss**» (Chaco)

■ **Vesres** : Creación de palabras por inversión de sílabas que se usa jergalmente o con fines humorísticos.

«Estoy en lo de villa mateando con él y jimmy. Pinta **sogui** abundante más tarde dijeron » (Corrientes)

«Uhhh me acuerdo si no habré saltado el muro del aguapey par colarme a los **cequin**. (cequín “fiesta de quince”)» (Chaco)

■ **Intejerccciones**

«**Aijué**, encima me decís vieja, re que no pinta esto facundo jaja ya te dije como es la onda, fin » (Formosa)

«**Ains**, una mujer hablando de fútbol.» (Formosa)

«Al fin una buena: hora libreeee! **Yirr** » (Corrientes)

■ **Guaranismos** Cabe destacar la detección de términos en guaraní en la región guaranítica¹. Un ejemplo de esto fueron las palabras *angá*, *angauí* y *mitai*. Si bien estas palabras provienen del guaraní, son utilizadas en oraciones en español. Como se puede ver en la tabla 4.2 el contraste entre las frecuencia normalizadas ² de la región

¹ Teniendo a las regiones dialectales marcadas por Vidal de Battini

² La frecuencia normalizada es una medida de estandarización que indica la cantidad de veces que aparece una determinada forma por cada millón de palabras.

	Región Guaranítica		Región Litoral	
	#Ocurrencias	Frecuencia Normalizada	#Ocurrencias	Frecuencia Normalizada
Angá	548	45,03	6	0,21
Angaú	205	16,84	0	0
Mitai	175	15,69	1	0,036

Tab. 4.2: Cantidad de ocurrencias y frecuencias normalizadas de las palabras en la región guaranítica y la del litoral. La cantidad total de palabras en la región guaranítica es de 12.167.635, mientras que la cantidad de términos en la región litoral es 27.477.861

guaranítica y la del litoral da una noción de la importancia que tienen estos términos en norte argentino.

Estos términos serán agregados al diccionario del habla de los argentinos [dL08].

4.5. Validación estadística

4.5.1. Test Hipergeométrico

Luego de realizar el listado de palabras ordenado por el valor de la información, se aplicó un test estadístico para tener mayor confianza de que las palabras clasificadas como contrastivas realmente tienen esta propiedad y no fueron producto del azar. Se seleccionó un conjunto de palabras significativas a nivel lingüístico a partir de las 5000 palabras consideradas más contrastivas por nuestra métrica.

Decidimos elegir el test hipergeométrico ya que queremos ver que la palabra sobre la que se hace el test no estuvo sobrerepresentada en comparación con la población. Asumimos que la cantidad de ocurrencias de una palabra se puede modelar con una distribución hipergeométrica ya que se puede pensar como un experimento donde se obtuvieron k palabras exitosas en una región con n palabras y un total de N palabras en la Argentina. Las regiones que utilizamos para cada palabra son el conjunto de provincias que cubren el 80 % de las ocurrencias de dicho término. Luego, queremos calcular la significancia estadística de haber obtenido esas k palabras exitosas.

Luego, por cada palabra seleccionada como contrastiva le aplicamos el test estadístico con la siguiente hipótesis nula: la palabra tienen un uso homogéneo en las distintas regiones de la Argentina, es decir que la frecuencia de ocurrencias de cada palabra debería ser similar independientemente de la región. Por lo tanto, en caso de que la palabra sea contrastiva deberíamos obtener una baja probabilidad de haber obtenido diferencias entre las frecuencias de la palabra en una región con el resto del país. Para aplicar el test hipergeométrico representamos los datos sobre la palabra en una tabla de 2x2 como la de la tabla 4.3.

	#Palabras Sobre Region	#Palabras en el resto de Argentina	Total
# Palabras w	k	K-k	K
# Palabras \neq w	n-k	N + k - n - K	N - K
Total	n	N - n	N

Tab. 4.3: Tabla de contingencia

En primer lugar hicimos el test estadístico sobre las palabras del conjunto de datos de

desarrollo para ver resultados preliminares. El test lo realizamos sobre palabras candidatas a ser contrastivas identificadas a través de nuestra métrica. Una vez realizado este test obtuvimos los p-valores de la figura 4.1. Debido a que realizamos múltiples test tuvimos que aplicarle una corrección para evitar falsos positivos. Decidimos utilizar la corrección de Benjamini–Hochberg [BH95] con $\alpha = 0,5$.

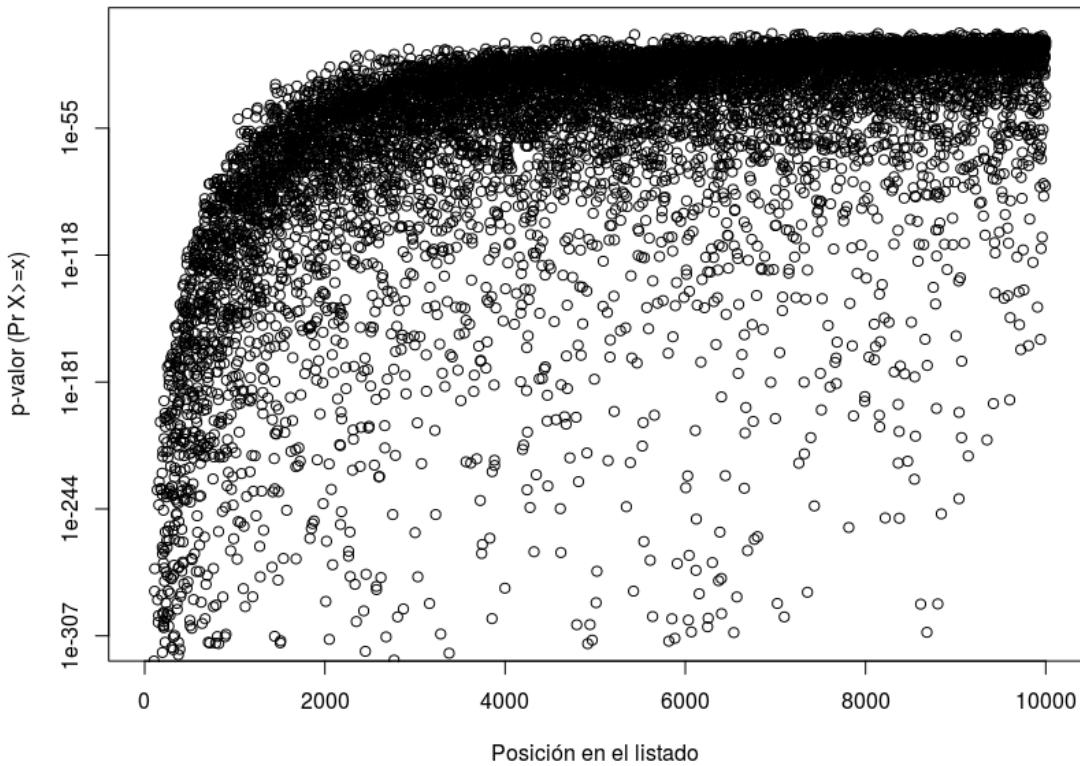


Fig. 4.1: Gráfico de dispersión de los p-valores del test hipergeométrico sobre las primeras 10000 palabras

Ante los p-valores tan bajos, decidimos hacer el test estadístico sobre palabras que consideramos que no deberían tener una frecuencia muy variada en las distintas regiones. Realizamos el test para las palabras {que, cuando, hola} y estos también dieron p-valores menores a 0.001. Frente a esta situación investigamos las posibles causas de este fenómeno.

Un estudio de Adam Kilgariff titulado *Language is never, ever, ever, random* [Kil05] analiza el uso de χ^2 y log-likelihood test son problemáticos, ya que se basan en la suposición de que todas las muestras son estadísticamente independientes. El autor afirma que debido a que el lenguaje no es aleatorio (ya que hablamos y escribimos con una intencionalidad) y que la hipótesis nula de los test estadísticos asumen aleatoriedad, cuando los datos provienen de fenómenos lingüísticos sobre corpus la hipótesis nula nunca es cierta. Kilgariff agrega que cuando hay suficientes datos (casi) siempre podemos ser capaces de rechazar la hipótesis nula. Con esto, el autor no quiere decir que los tests estadísticos no sirven, incluso menciona que en estadística se realizan tests que suponen la aleatoriedad

cuando el fenómeno a estudiar no presenta esa propiedad. El problema está en realizar estos tests cuando el fenómeno es tan arbitrario como el lenguaje y de esta manera se simplifica en demasiado la hipótesis de aleatoriedad. Kilgariff también hace un análisis sobre trabajos previos y comenta el caso en el que un estudio quería encontrar palabras con frecuencias significativas entre el inglés británico y el americano, representados por el Corpus Brown(inglés americano) y el Lancaster-Oslo-Bergen Corpus(inglés británico). Para cada palabra testearon la hipótesis nula la cual afirmaba que la diferencia entre las frecuencias sobre los dos corpus se debía a una fluctuación aleatoria. El test lo hicieron a partir de muestras obtenidas aleatoriamente de los corpus mencionados. En este estudio se marcaron las palabras donde la hipótesis nula se rechazaba con distintos niveles de confianza. Las listas sugieren que la mayor parte de las palabras *comunes* fueron marcadas, es decir que el test estadístico sugería que todas estas palabras tenían diferencias significativas en su uso. Esto que comenta Kilgariff es lo que tuvimos como resultado a partir de nuestro test hipergeométrico. Lo interesante es que el autor atribuye ese fenómeno a la esencia no aleatoria del lenguaje. Si bien sabíamos que al realizar el test hipergeométrico asumíamos que todas las palabras son estadísticamente independientes, pensamos que esta suposición no iba a afectar tanto los resultados como lo hizo.

Tanto Kilgariff como Paquot y Bestgen afirman que es posible representar los datos de manera diferente y hacer uso de otros tests asumiendo la independencia entre los textos y no entre las palabras. De esta manera se pueden observar la distribución de las palabras dentro de un corpus. En un estudio más reciente, Lijffijt et al. [LNS⁺] analiza algunas de estas alternativas para hacer un test de hipótesis en los cuales no se asume el modelo de *bag-of-words*. Este modelo es el que calcula las frecuencias de las palabras sobre todos los textos, sin tener en cuenta la distribución de las frecuencias en cada uno. En ese trabajo se analizaron el test t de Welch, el test de los rangos con signo de Wilcoxon(Wilcoxon rank sum), el de Bootrstrap y el de tiempo entre llegadas (inter-arrival time). Lijffijt explica que la diferencia entre estos tests con los que asumen el modelo de *bag-of-words*(χ^2 y log-likelihood test) reside en la representación de los datos, ergo la unidad de observación:

Para los tests que suponen los modelos de bolsa de palabras o *bag-of-words*, los datos se representan en una tabla de contingencia de 2x2 y el número de muestras equivale al número de palabras en el corpus, mientras que en los otros cuatro tests, los datos son representados en una lista de frecuencias o una lista de *tiempos de llegada*. En estos casos, el número de muestras es mucho menor que la cantidad de palabras en el corpus. Se destaca el número de muestras de los modelos ya que este número generalmente determina nuestro nivel de seguridad en relación a los valores estimados. Es por esto que los resultados experimentales muestran que los tests de modelos de *bag-of-words* tienen una excesiva alta confianza en los valores estimados de la frecuencia media de las palabras, en el contexto de comparación estadística entre dos corpus. [LNS⁺].

4.5.2. Test t de Welch

Basándonos en las propuestas de Lijffijt, decidimos utilizar el test de Welch. Este nos provee un valor de probabilidad para rechazar la hipótesis nula la cual afirma que las medias de las dos distribuciones son iguales. Sean S y T dos corpus y sea q la palabra sobre la cual se va a hacer el test, sea x_1 la media de la frecuencia de la palabra q sobre los textos de S , y sea s_1 la desviación estándar. Análogamente, sea x_2 la media de la frecuencia de q en los textos T y s_2 la desviación estándar. El estadístico t se calcula con la ecuación 4.1. Las suposiciones del test consisten en que todos los textos son estadísticamente independientes

y que la media de las frecuencias proviene de una distribución normal. En nuestro caso, agrupamos todos los tuits de cada usuario representando un texto. De esta manera, cada provincia tiene alrededor de 900 textos formados por distintos usuarios³. Luego, el test es aplicado a cada palabra con las frecuencias entre dos corpus: uno está formado por todos los textos de los usuarios que provienen de las provincias en donde se cubre el 80 % de las ocurrencias, el otro consiste en los textos creados por usuarios del resto de las provincias. Notar que asumir la independencia entre todos los textos de un usuario de una provincia, es una suposición más débil que la de asumir independencia en todos los textos generados en una provincia como se hacía en el modelo de bolsa de palabras.

$$t = \frac{x_1 - x_2}{\sqrt{\frac{s_1^2}{|S|} + \frac{s_2^2}{|T|}}} \quad (4.1)$$

Realizando el test de Welch no queda claro si la distribución de los p-valores es la esperada. Ante esta posibilidad conjeturamos con que quizás la comparación entre en la hipótesis que formulamos en el análisis estadístico y la contrastividad dada por la métrica no es la más adecuada. Igualmente, no se replicó de manera idéntica el trabajo de Lifjijft et.al ya que nosotros cambiamos de corpus por cada palabra. La causa de esto es la que definimos una región compuesta de provincias en función de las frecuencias de la palabra en las distintas provincias argentinas. El análisis de significancia a partir de test estadísticos sobre fenómenos lingüísticos de corpus no es algo que este resuelto en la actualidad. Dejamos como trabajo futuro el cálculo de un análisis estadístico comparable a nuestra métrica ya que está por fuera del alcance de esta tesis.

Sin embargo, a partir del test de Welch obtuvimos algunos resultados que dan un indicio de las virtudes de la métrica desarrollada, los cuales exponemos a continuación. Uno de ellos es la comparación de las tasas de rechazo de la hipótesis nula sobre las tres métricas desarrolladas, I_u , I_w y I mencionadas en 3.8, 3.7 y 3.12 respectivamente. Calculamos la taza de rechazo para las palabras en distintos intervalos del listado ordenado según las tres métricas elegidas: la métrica que tiene en cuenta la entropía de palabras, la valora la entropía de personas, y la que contiene a los dos factores. En la figura 4.2 se muestran los resultados. La métrica elegida, la cual tiene a ambos factores en consideración tiene una mejor taza de rechazo de la hipótesis nula en las palabras consideradas contrastivas y una menor taza de rechazo para el resto. Es importante notar que el test de Welch que realizamos tiene como muestras las distintas frecuencias de todos los usuarios en cada región. Por lo tanto es razonable obtener un resultado como este, en el cual las métricas que consideran la dispersión de las frecuencias sobre todos los usuarios para distinguir el nivel de contrastividad de la palabra.

También es importante destacar que a medida que uno se aleja de las palabras más contrastivas de acuerdo a nuestra métrica, la tasa de rechaza es menor. Esto refleja el buen comportamiento de la métrica. Este resultado se puede observar también en la figura 4.3 donde se detalla la distribución de los p-valores en el conjunto de las primeras 5000 palabras y el resto de los términos del listado.

³ La cantidad de usuarios recolectados por cada provincia se encuentra detallada en la tabla 2.1

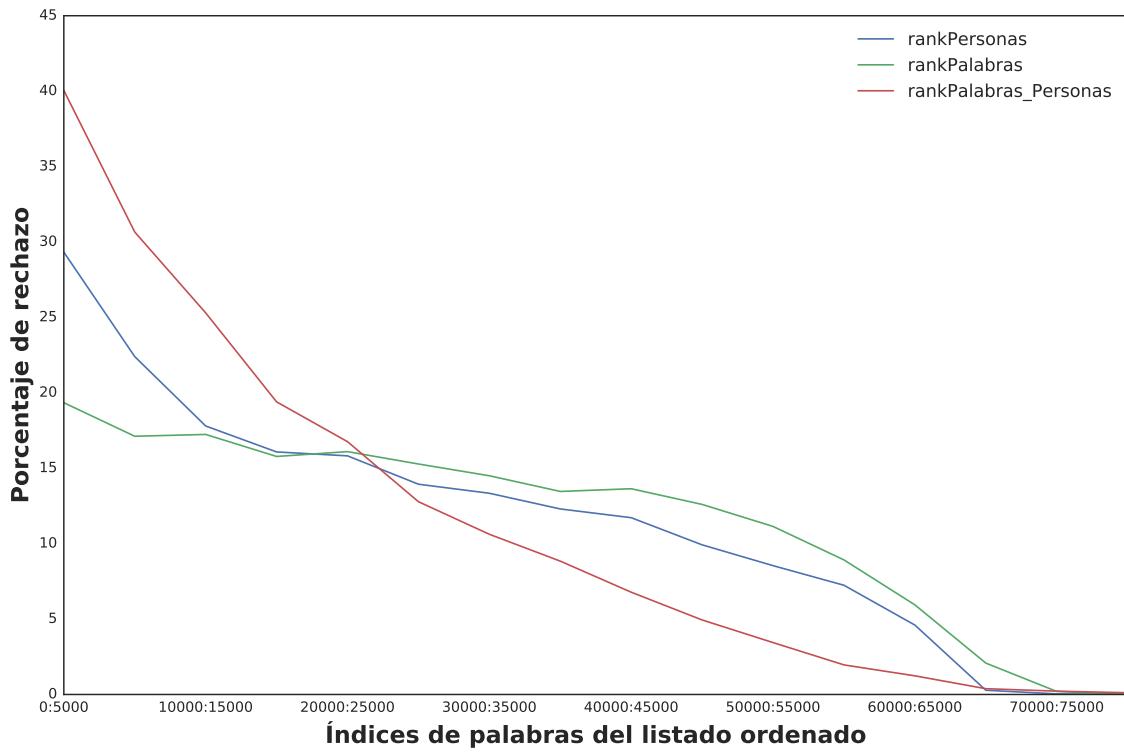


Fig. 4.2: Resume la tasa de rechazo de la hipótesis nula en los distintos conjuntos de palabras, los cuales varían según el índice de estas en el listado ordenado según la métrica elegida.

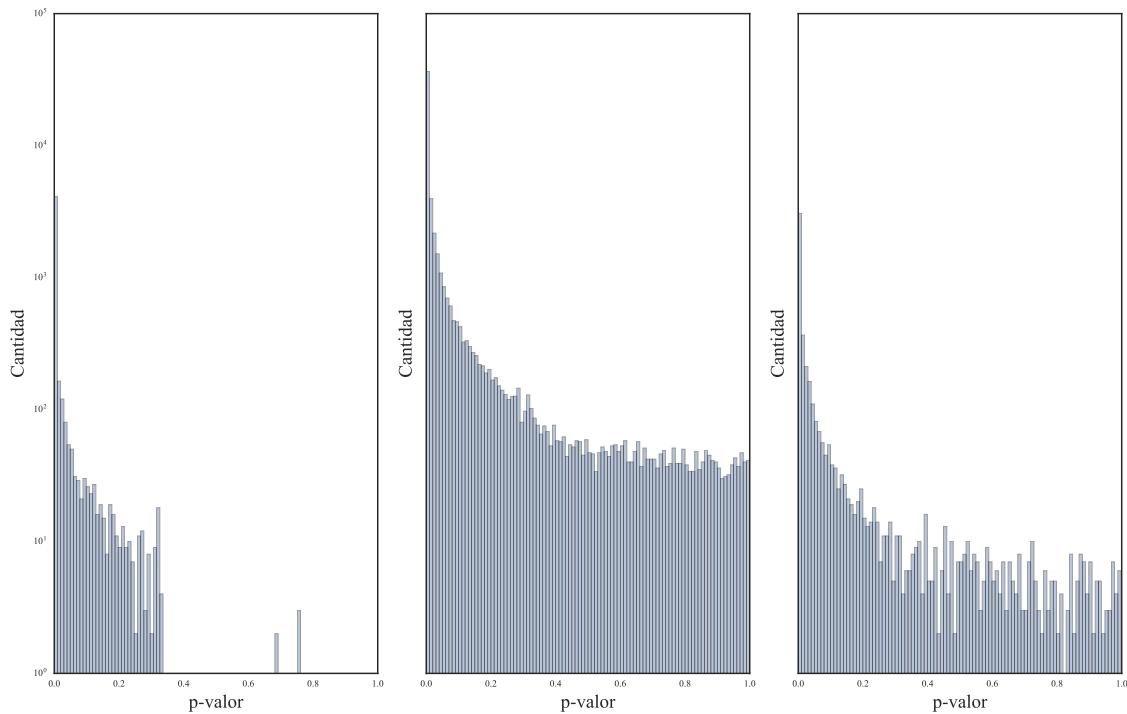


Fig. 4.3: Distribución de p-valores (sin corrección de tests múltiples) en las primeras 5000 palabras del listado y el resto de las palabras.

5. CONCLUSIONES Y TRABAJO FUTURO

5.1. Conclusiones y trabajo futuro

En el presente trabajo desarrollamos una métrica que de un índice contrastivo del uso de una palabra en distintas regiones. Para usar esta métrica recolectamos un conjunto de datos de textos de la Argentina a través de la API de Twitter.

La métrica que realizamos usa la entropía para medir la variación de la cantidad de ocurrencias y de la cantidad de usuarios que la utilizaban en las diferentes provincias del país. Creamos un listado de palabras ordenado según el valor de la información calculada. Sobre este listado, se seleccionaron las 5000 palabras con mayor valor de contrastividad para realizar una validación lingüística por parte de la Academia Argentina de Letras. La validación arrojó un resultado con alrededor de 300 palabras dignas de estudio por cada 5000 palabras, es decir, 1 palabra cada 17. A pesar de que no existen otros proyectos que provean un término de comparación para evaluar el grado de éxito implicado en esta relación, no cabe ninguna duda de que, al menos en la detección de coloquialismos locales actualmente en uso, la herramienta plantea un verdadero punto de inflexión para la lexicografía contrastiva. Varias de las palabras detectadas a partir de la métrica desarrollada serán agregadas al Diccionario del Habla de los Argentinos.

En este trabajo se analizan las regiones formadas con una provincia como unidad regional, sin embargo se puede cambiar esta unidad para replicar el análisis con distinta granularidad. De esta manera se podría ver las palabras contrastivas en los distintos países hispanoparlantes y comparar las variaciones entre regiones más grandes.

Uno de los desafíos que quedan para hacer es el de clasificar las regiones en clusters, obteniendo así un indicio de las regiones dialectales actuales. De esta manera se podría considerar la vigencia de las regiones propuestas por Vidal de Battini.

También, el proceso de normalización se podría mejorar para tener una mayor precisión de las palabras utilizadas. También se podría agregar un sistema de reconocimiento de nombres de entidades para destacar también ciertos nombres propios de manera tal que el listado de palabras tenga más alertas sobre términos sin interés lingüístico.

Finalmente se puede hacer un análisis sintáctico de las oraciones, y un estudio de contrastividad comparando la distribución de los n-gramas, a diferencia del análisis por palabra hecho en este trabajo.

Es importante señalar las ventajas de *Twitter* ya que nos permitió recolectar un volumen grande de datos de texto, escritos por distintas personas con información de su localización. Acerca de las desventajas de esta plataforma podemos los errores ortográficos de los textos, los cuales contienen abreviaciones o cambios para lograr un énfasis en el discurso. Todo esto conlleva a un aumento de la dificultad para normalizar el texto. Creemos sin embargo, que el flujo de datos prevalece a la hora de decidir una palataforma para recolectarlos.

6. APÉNDICE

6.1. La entropía como medida del desorden

Para ver la cantidad de información que nos aporta cada palabra se hará una introducción a la teoría de la información, específicamente los conceptos que introdujo Claude Shannon[Sha01]. Para entender estos conceptos es útil tener una descripción matemática del mecanismo que genera la información. Para eso se define a la *fuente* que emite señales de un alfabeto $S = \{s_1, s_2, \dots, s_q\}$ de acuerdo a una función de probabilidad fija. Si la fuente emite señales estadísticamente independientes decimos que es una *fuente de memoria nula* y un símbolo s está completamente determinado por el alfabeto S y las probabilidades: $P(s_1) P(s_2) \dots P(s_q)$

Sea X una variable aleatoria discreta con posibles valores $\{x_1, x_2, \dots, x_q\}$ y una función de probabilidad $P(X)$, luego: $H(X) = E[I(X)] = E[-\log(P(X))]$. donde X es una variable aleatoria con posibles valores $\{x_1, \dots, x_n\}$ y P es una función de probabilidad.

Los símbolos con menor probabilidad son los que aportan más información. Esto va de la mano con nuestra intuición ya que si entendemos a los símbolos como palabras de un texto, las palabras más utilizadas como *de* o *que* aportan menos información que la palabra *celular*. Observaciones:

- La entropía es máxima cuando los eventos de X son equiprobables. En este caso, si hay n eventos con una probabilidad de $\frac{1}{n}$ cada uno, el valor de la entropía es de $\log n$.
- La entropía es 0 si y solo si todas las probabilidades son 0 a excepción de una con probabilidad igual a la unidad.

Dado que la entropía es máxima cuando los eventos de X son equiprobables, se suele decir que es una medida del desorden.

6.2. Tablas y Gráficos

Conjunto de Provincias			
Salta-Jujuy	Mendoza-San Juan	Chubut-Santa Cruz-T. Del Fuego	Chaco-Corrientes-Formosa
tribuno	mza	austral	anga
salteño	zonda	chilote	teresss
orán	secamente	calafa	músicas
tartagal	sanjuaninos	chilota	argela
salteña	sanjuanino	palmaso	angaaa
martearena	asar	vueltines	olo
oran	tomba	riviera	cuchale
yuto	queras		corrientesss
purmamarca	traica		angá
yutos	sopaipillas		cts
gorriti	ardente		correntinas
quijsano	secamente		iburrr
tabacal	jáchal		cheraa
desentierro	virreina		cheraá
huaico	tombaaa		bofill
pichanal	mansooo		receppp
diableros	tombino		
bandy	parisi		
aramayo	asadaaaa		
ñáñoo			
colque			
urkupiña			
juy			
guachipas			

Tab. 6.1: Palabras cubiertas sobre el 80 % de las ocurrencias totales por el conjunto de provincias a partir de las 5000 palabras más contrastivas (de acuerdo al valor de la información).

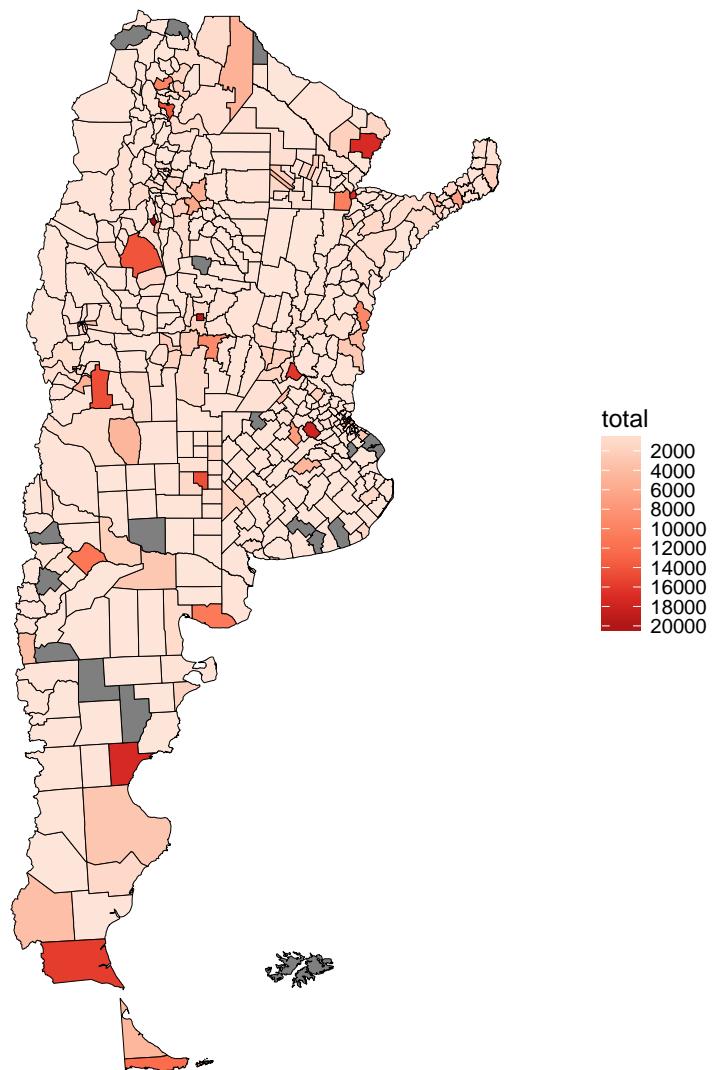


Fig. 6.1: Mapa con la distribución de los tuits que incluyeron sus coordenadas geográficas.

7. BIBLIOGRAFÍA

Bibliografía

- [AV95] Manuel Almeida and Carmelo Vidal. Variación socioestilística del léxico: un estudio contrastivo. *Boletín de filología*, 35(1):Pág–50, 1995.
- [Ávi04] Raúl Ávila. ¿ el fin de los diccionarios diferenciales? ¿ el principio de los diccionarios integrales? 2004.
- [Baa01] R Harald Baayen. *Word frequency distributions*, volume 18. Springer Science & Business Media, 2001.
- [BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.
- [Bis06] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- [dL08] Academia Argentina de Letras. *Diccionario del habla de los argentinos*. Emecé Editores, 2008.
- [Doy14] Gabriel Doyle. Mapping dialectal variation by querying social media. In *EACL*, pages 98–106, 2014.
- [Eis14] Jacob Eisenstein. Identifying regional dialects in online social media. Georgia Institute of Technology, 2014.
- [EOSX10] Jacob Eisenstein, Brendan O’Connor, Noah A Smith, and Eric P Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1277–1287. Association for Computational Linguistics, 2010.
- [Esp] Real Academia Española. Banco de datos (corpes xxi)[en línea]. *Corpus del español del siglo XXI (CORPES)*.
- [GS14] Bruno Gonçalves and David Sánchez. Crowdsourcing dialect characterization through twitter. *PloS one*, 9(11):e112074, 2014.
- [KG03] Adam Kilgarriff and Gregory Grefenstette. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347, 2003.
- [Kil05] Adam Kilgarriff. Language is never, ever, ever, random. *Corpus linguistics and linguistic theory*, 1(2):263–276, 2005.
- [Liu12] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.
- [LNS⁺] Jefrey Lijffijt, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. Significance testing of word frequencies in corpora.

-
- [MH11] Tony McEnery and Andrew Hardie. *Corpus linguistics: Method, theory and practice*. Cambridge University Press, 2011.
 - [Mon01] Marcelo A Montemurro. Beyond the zipf–mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, 300(3):567–578, 2001.
 - [MZ10] Marcelo A Montemurro and Damián H Zanette. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153, 2010.
 - [PD11] Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. *Icwsrm*, 20:265–272, 2011.
 - [PP10] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, 2010.
 - [Sha01] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.
 - [SOM10] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.
 - [TSSW10] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Icwsrm*, 10(1):178–185, 2010.
 - [Zim06] Klaus Zimmermann. El fin de los diccionarios de mexicanismos, colombianismos, argentinismos, cubanismos etc. la situación de la lexicografía del español de américa después de la publicación de los diccionarios contrastivos del español de américa: Español de amér. *Estudios de lingüística del español*, 23, 2006.
 - [Zip16] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.